# Discovering and Mitigating Indirect Bias in Attention-Based Model Explanations

**Farsheed Haque**
University of North Carolina
at Charlotte
fhaque@uncc.edu

**Depeng Xu**
University of North Carolina
at Charlotte
dxu7@uncc.edu

**Shuhan Yuan**
Utah State University
shuhan.yuan@usu.edu

## Abstract

As the field of Natural Language Processing (NLP) increasingly adopts transformer-based models, the issue of bias becomes more pronounced. Such bias, manifesting through stereotypes and discriminatory practices, can disadvantage certain groups. Our study focuses on direct and indirect bias in the model explanations, where the model makes predictions relying heavily on identity tokens or associated contexts. We present a novel analysis of bias in model explanation, especially the subtle indirect bias, underlining the limitations of traditional fairness metrics. We first define direct and indirect bias in model explanations, which is complementary to fairness in predictions. We then develop an indirect bias discovery algorithm for quantitatively evaluating indirect bias in transformer models using their in-built self-attention matrix. We also propose an indirect bias mitigation algorithm to ensure fairness in transformer models by leveraging attention explanations. Our evaluation shows the significance of indirect bias and the effectiveness of our indirect bias discovery and mitigation.

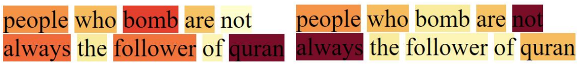
## 1 Introduction

Discrimination is the unfair treatment or prejudice directed towards individuals, groups, or certain ideas or beliefs, intentionally or unintentionally. It frequently entails making stereotypes about others and acting in a manner that disadvantages one group while favoring another (Webster et al., 2022). The pervasive nature of bias extends to machine learning, prominently manifesting in the domain of Natural Language Processing (NLP) (Bansal, 2022). As NLP becomes increasingly integral to everyday life, largely due to the advancements brought by the transformer-based models (Wolf et al., 2020; Dai et al., 2019), addressing fairness in this field is of utmost importance.

In recent years, NLP researchers have undertaken efforts to identify and mitigate discrimination against specific groups, such as gender (Thelwall, 2018), race (Kiritchenko and Mohammad, 2018), age (Diaz et al., 2018), religion (Bhatt et al., 2022), disability (Venkit and Wilson, 2021), etc. They focus on the model's tendency to exploit spurious correlations (Liusie et al., 2022; Wang et al., 2022) between the predicted label and explicit words linked to certain protected attributes, such as "he", "she", "Alice", "Bob", "Russian", "Muslim", etc. For instance, in a hate speech detection task, an unfair transformer-based model would see the word "Muslim" (also a protected attribute) in a sentence and classify it as hate speech instantly by assigning high attention to the word "Muslim", rather than understanding the whole message of the sentence. This is referred to as the legal concept of disparate treatment (Supreme Court of the United States, 1971), that is the outcomes have intended direct discrimination due to choices made explicitly based on membership in a protected class. The existing methods can only handle discriminatory cases where there is a representative token present in the text directly associated with the protected group, e.g., token "Muslim" for the Islam religion. It also requires the NLP practitioners to manage a pre-determined list of candidate tokens.

In contrast to disparate treatment, disparate impact (Supreme Court of the United States, 1971) is the legal theory that outcomes should not be different based on individuals' protected class membership, even if the process used to determine that outcome does not explicitly base the decision on that membership but rather on proxy attributes. Even without the presence of any direct indicating token in the text, the model still excessively relies on context learned from biased training data, which results in unintended subtle indirect discrimination in the prediction. Such indirect association is case by case. It is difficult to pre-determine a candidate token list. Remarkably, no prior studies have explicitly delved into indirect discrimination in NLP,

*Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1599–1614
June 16-21, 2024 ©2024 Association for Computational Linguistics


people who bomb are not always the follower of quran
people who bomb are not always the follower of quran

(a) Biased Model (b) Unbiased Model

Figure 1: An example of token-wise model explanation. The darker color indicates a higher importance.

to the best of our knowledge.

In this work, we want to bridge the gap between disparate treatment and disparate impact in NLP models. The black-box deep learning models tend to over-learn the biased data during training, which results in shortcuts in decision-making without valid explanations. Figure 1 illustrates how a model trained to mitigate direct bias against Islam religion through "Muslim" still falsely categorizes a statement as hate speech because the model's attention is biased emphasized on the sensitive context like the word "quran". An unbiased model would make a negative prediction based on "not always". To investigate bias in the model's local explanations, we first define direct and indirect bias (in Section 4). They complement the traditional outcome-association-based group fairness notions, such as demographic parity and equal opportunity. We then propose a novel bias discovery method to evaluate transformer-based models on disparate impact (in Section 5). It leverages a secondary transformer-based model dedicated to classifying the protected attribute from the association presented in the training data. We compare the faithful explanations of the primary, potentially biased model, with those of this secondary model. By examining the similarity between their decision-making patterns, we quantify indirect bias through a new proposed metric called the **Area Under the Similarity Curve (AUSC)**. Furthermore, we then proceed to mitigate the detected indirect bias through a similarity-based constraint, which is coupled with mitigating direct bias through data Resampling and adversarial learning (in Section 6). In our experiment, we show the significance of indirect bias, the effectiveness of our indirect bias discovery and mitigation algorithms, and the advantage of mitigating indirect bias in model explanations (in Section 7). Thus, our primary contributions are threefold: (1) we establish the problem of fairness in model explanations by formally defining direct and indirect bias; (2) we propose an **Indirect Bias Discovery (IBD)** framework tailored to quantitatively evaluate indirect bias in transformer models; and (3) we develop a novel **Indirect Bias Mitigation (IBM)** algorithm that ensures fairness using model explanations. Our codes are available at https://github.com/FarsheedHaque/Indirect-Bias

## 2 Related Work

### 2.1 Bias and Mitigation

An increasing body of work has been conducted on direct bias discovery in NLP and ways to mitigate it. Researchers have focused on classification tasks and how societal biases (Hutchinson et al., 2020; Dinan et al., 2020; Xia et al., 2020), can impact a model's prediction. While these studies work on one type of social bias at a time others have tried to make a generalized method to quantify any sort of existing bias (Czarnowska et al., 2021). (Hovy and Prabhumoye, 2021), argues that these direct biases originate mainly from five sources. To observe bias (Bansal, 2022), talks about existing metrics in nlp.

Many attempts have been made to mitigate bias by solving sub-problems. Generally, all bias mitigation approaches fall under three categories (Mehrabi et al., 2021). **Pre-processing**, when mitigation happens before feeding the biased data into the model. (Kamiran and Calders, 2011) resamples the biased dataset to get an unbiased dataset. (Brunet et al., 2019) tries to locate the bias that exists in training data and remove it so that the model can train on unbiased data. However, the model has to allow such modification in the training data (Bellamy et al., 2018). **In-processing** mitigation is such, where the model's algorithm is modified to tackle bias while training on biased data. Adversarial learning (Zhang et al., 2018), is a prime example of in-process bias mitigation. Other solutions like causal mediation analysis (Vig et al., 2020), entropy-based attention regularization (Attanasio et al., 2022) are offered to mitigate bias using regularization terms and (He et al., 2022) uses a different model to predict the sensitive attribute to use their rationale to mitigate bias in the training time. Finally, **post-processing**, involves using a separate set of data, not used during the model's training, to evaluate the model after its training phase is complete (d'Alessandro et al., 2017). In (Bolukbasi et al., 2016), the author introduced an equalization process for every pair of gender-specific words to ensure fairness.

### 2.2 Attention Interpretation

Attention interpretability in NLP is crucial for understanding the biased decision-making process of transformer-based models (Mehrabi et al., 2022). Self-attention mechanisms are structured as multi-layered entities, with each layer encompassing multiple heads. Given the complexity of this high-dimensional architecture, it is a challenge to interpret the decision-making process of self-attention. As a remedy, researchers often project the self-attention representations into a more manageable lower-dimensional space (Mylonas et al., 2022). Several operations on heads and layers, such as averaging (Wang et al., 2019) and summation (Schwenke and Atzmueller, 2021), have been proposed to simplify this process. These operations inherently rank tokens by their significance by aggregating column-wise data into unified matrices for heads (Schwenke and Atzmueller, 2021; Mathew et al., 2021; Chefer et al., 2021). Multiplication is also a good layer operation (Chefer et al., 2021) because it can amplify the signals that might be muted using other techniques. The careful sequencing of these, among other operations, can be used to aggregate self-attention scores to achieve an interpretation.

While some scholars (Jain and Wallace, 2019; Pruthi et al., 2019) suggest that the attention mechanism may not serve as a dependable means for understanding how models make decisions, recent research indicates that methods for measuring faithfulness can effectively assess the utility of these interpretive approaches. A strong faithfulness score implies an effective attention aggregation technique, which in turn can provide reliable interpretations. (Mylonas et al., 2022) introduced Ranked Faithful Truthfulness, aimed specifically at evaluating methods of attention aggregation. Additionally, studies such as (DeYoung et al., 2020) have developed more generalized metrics, including comprehensiveness and sufficiency, to assess the rationale behind a model's decisions.

## 3 Preliminary

Given an input sequence $\boldsymbol{x}$ with a corresponding protected attribute $s$ and a class label $y$. $\boldsymbol{x}$ is an ordered sequence of tokens represented as $\boldsymbol{x} = \{t_i\}_{i=1}^{N}$ with $t_i$ denoting the $i$-th token in the sequence and $N$ is the length of $\boldsymbol{x}$. The protected attribute $s$ is the protected group of the person associated with the text. It can be the composer or recipient of the text, or the target whom the text comments on. The value of $s = u$ (e.g., gender is female) is sometimes already expressed in $\boldsymbol{x}$ as a sensitive token $\underline{s}$ (e.g., "she"), i.e., $\underline{s} \in \boldsymbol{x}$, which is mostly studied by previous works. In this work, we do not require the presence of $\underline{s}$ in $\boldsymbol{x}$, where the protected attribute $s$ is a hidden context. The class label $y$ is the prediction target. A text classification model $f : \boldsymbol{x} \rightarrow y$ is trained on labeled text data $(\boldsymbol{x}, y)$. The model prediction for a sequence $\boldsymbol{x}$ is denoted as $\hat{y} = f(\boldsymbol{x})$. Specifically, we consider a state-of-the-art transformer-based model.

### 3.1 Prediction Outcome Fairness

**Demographic parity** is a notion of group fairness, where the model prediction is fair w.r.t. the values of protected attribute $s$ if $\hat{y}$ and $s$ are independent of each other (Zhang et al., 2018).

$$P(\hat{y} = 1 | s = u) = P(\hat{y} = 1 | s = v)$$

**Equality of Opportunity** is another notion of group fairness, where a model's predictions are deemed fair w.r.t. a protected attribute $s$ if the true positive rate of $\hat{y}$ is the same across different groups defined by $s$ (Zhang et al., 2018).

$$P(\hat{y} = 1 | s = u, y = 1) = P(\hat{y} = 1 | s = v, y = 1)$$

### 3.2 Self-Attention

When $f$ is a transformer-based model, the self-attention mechanism in $f$ plays a crucial role in understanding token relationships within the sequence $\boldsymbol{x}$. For each self-attention layer, the initial input is an $(N \times E)$ matrix where $N$ is sequence length and $E$ is embedding size. This matrix undergoes linear transformations to produce matrices $Q$(query), $K$(key), and $V$(value) of the same size.

$$A = softmax\left(\frac{Q.K^T}{\sqrt{E}}\right)V, \qquad (1)$$

where the dot product between $Q$ and $K$ is computed, and the result is scaled by dividing it by $\sqrt{E}$. The output undergoes a softmax function, resulting in $(N \times N)$ matrix called $A$ (Vaswani et al., 2017). This matrix encapsulates the attention-based relationships of every token $t_i$ in the sequence $\boldsymbol{x}$ to every other token.

In the classification task, certain tokens play a vital role in predicting $y$, and these tokens get high self-attention scores (Letarte et al., 2018). Let $\boldsymbol{t}^y$ denote the set of these ground-truth centric tokens where $\boldsymbol{t}^y \in \boldsymbol{x}$. The attention score of tokens in this set, represented as $A[\boldsymbol{t}^y]$ is notably high. The aggregated token-wise attentions often serve as local model explanations, which in return help to identify these ground-truth centric tokens $\boldsymbol{t}^y$.

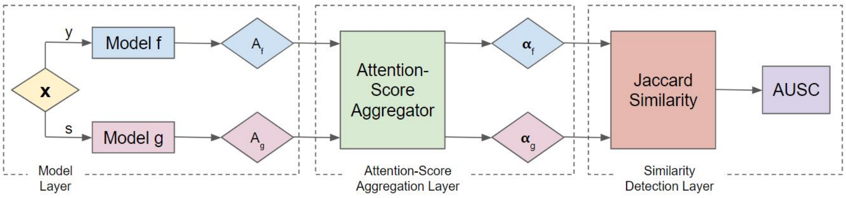

y
Model f
A_f
x
s
Model g
A_g
Model Layer
Attention-Score Aggregator
α_f
α_g
Attention-Score Aggregation Layer
Jaccard Similarity
AUSC
Similarity Detection Layer

Figure 2: Indirect Bias Discovery (IBD) Architecture

## 4 Direct and Indirect Bias

Consider a text classification model $f : \boldsymbol{x} \rightarrow y$ that is trained on labeled text data $(\boldsymbol{x}, y)$. There also exists a protected attribute associated with $\boldsymbol{x}$, which may or not be present in the text in the form of an identity token. Regardless of the bias in training data, it is essential to make sure the prediction $\hat{y}$ made by the trained model $f$ is unbiased w.r.t. $s$ not only in the predicted outcomes but also in the local explanations to justify the prediction. In this section, we formally define direct and indirect bias in the model explanations and therefore formulate new fairness notions.

**Direct Bias.** In text data, the protected attribute is sometimes (but not always) already present in the text sequence, i.e., $\underline{s} \in \boldsymbol{x}$. If a model explicitly makes predictions based on the sensitive token $\underline{s}$, we define such bias in the model explanations as direct bias. For a model $f$ with direct bias, the sensitive token $\underline{s}$ is among the key tokens for the model decision, i.e., $\underline{s} \in \boldsymbol{t}^y$, where $\boldsymbol{t}^y$ denotes the set of important tokens which $f$ makes the prediction $\hat{y}$ based on. The key token set $\boldsymbol{t}^y$ serves as the deciding factor in the model's local explanation.

**Theorem 1** *A model $f$ satisfies no direct bias in explanations if the sensitive token $\underline{s}$ is not explicitly used for model decisions, i.e., $\underline{s} \notin \boldsymbol{t}^y$.*

**Indirect Bias.** Other than the sensitive token $\underline{s}$, when the model makes a prediction, it can also over-exploit context $\boldsymbol{t}^s$ in the text which is highly correlated to $s$. We define such bias in the model as indirect bias. For a model with indirect bias, a subset of the sensitive context tokens $\boldsymbol{t}^s$ is among the key decision-making tokens $\boldsymbol{t}^y$, i.e., $\boldsymbol{t}^s \cap \boldsymbol{t}^y \neq \emptyset$.

**Theorem 2** *A model $f$ satisfies no indirect bias in explanations if the sensitive context tokens are not used for model decisions, i.e., $\boldsymbol{t}^s \cap \boldsymbol{t}^y = \emptyset$.*

## 5 Indirect Bias Discovery (IBD)

Direct and indirect bias evaluate a model's fairness in terms of its decision-making process, a.k.a. model explanations. An unbiased transformer-based model pays high attention to the set of these ground-truth centric tokens $\boldsymbol{t}^y$, whereas a model with indirect bias pays high attention to a set of tokens $\boldsymbol{t}^s$ that is associated with $s$. In practice, either $\boldsymbol{t}^y$ or $\boldsymbol{t}^s$ is not annotated in the text. A model $f$ can provide local explanations in the form of $\boldsymbol{t}^y$. The key challenge to examine indirect bias is to identify $\boldsymbol{t}^s$. To separate $\boldsymbol{t}^s$ from $\boldsymbol{t}^y$ and to discover indirect bias in model $f$ we propose an Indirect Bias Discovery (IBD) architecture. Figure 2 shows a general overview of our proposed architecture. It is divided into three components - model layer, attention-score aggregation layer, and similarity detection layer.

**Model Layer** is used to fine-tune our target model $f$ on sequence $\boldsymbol{x}$. The goal of this fine-tuned $f$ is to successfully predict $\hat{y}$ where $\hat{y} = f(\boldsymbol{x})$. We also get the attention-score matrix $A_f[\{t_i\}_{i=1}^N]$ for $\boldsymbol{x}$ in model layer which we can use to identify $\boldsymbol{t}^y$ later. This layer also has another helper model $g$ fine-tuned to predict the protected attribute $s$ of $\boldsymbol{x}$ such that $\hat{s} = g(\boldsymbol{x})$. Model $g$ also gives us the attention-score matrix $A_g[\{t_i\}_{i=1}^N]$ for $\boldsymbol{x}$ which we can use to identify $\boldsymbol{t}^s$ later. Then, $A_f$ and $A_g$ are fed into the next layer as inputs to get the interpretation of the decision-making process of model $f$ and $g$ respectively.

**Attention-Score Aggregation Layer** takes high-dimensional matrices, $A_f$ and $A_g$ and maps them into one-dimensional vectors, $\alpha_f$ and $\alpha_g$. These vectors encapsulate the importance scores for the token set $\{t_i\}_{i=1}^N$ originating from $A_f$ and $A_g$, respectively. To achieve this we devised a self-attention score aggregator using different combinations of summation, multiplication, average, and maximum. From different combinations, we took one that performs best on the faithfulness metrics of comprehensiveness and sufficiency (DeYoung et al., 2020). Sufficiency evaluates how sufficient an aggregation is for making a prediction, while comprehensiveness assesses if all the selected elements are essential for the prediction. A minimal reduction in sufficiency and a significant drop in comprehensiveness suggest a high level of faithfulness. Our attention-score aggregator follows the operations as in Algorithm 1 below.

**Algorithm 1** Faithful Attention Aggregator

```
Require: model $f$, input instance $x$
$L, H \leftarrow$ number of layers and heads in $f$
$A[l][h] \leftarrow$ attention matrix of layer $l$, head $h$ in $f$ given $x$
for each combination of $head\_op$, $layer\_op$, $token\_op$ in $[sum, mul, mean, max]$ do
    $B = head\_op(A[l])$
    $C = layer\_op(B)$
    $\alpha = token\_op(C)$
    Evaluate the faithfulness of $\alpha$
end for
return Best aggregation combination based on faithfulness metrics and the corresponding $\alpha$
```

**Similarity Detection Layer** finds the $\boldsymbol{t}^y$ and $\boldsymbol{t}^s$ to detect indirect bias in model $f$. To achieve this, the layer takes $\alpha_f$ and $\alpha_g$ as inputs. A subset $\boldsymbol{t}_f^k$ is selected from $\boldsymbol{x}$, which comprises the top $k\%$ importance scores in $\alpha_f$. $\boldsymbol{t}_f^k$ is a hypothesis of $\boldsymbol{t}^y$ based on $f$. Consequently, a subset $\boldsymbol{t}_g^k$ is selected from $\boldsymbol{x}$, which comprises the top $k\%$ importance scores in $\alpha_g$. $\boldsymbol{t}_g^k$ is a hypothesis of $\boldsymbol{t}^s$ based on $g$. The similarity between the subsets $\boldsymbol{t}_f^k$ and $\boldsymbol{t}_g^k$ is calculated as below.

$$\phi = J(\boldsymbol{t}_f^k, \boldsymbol{t}_g^k) = \frac{|\boldsymbol{t}_f^k \cap \boldsymbol{t}_g^k|}{|\boldsymbol{t}_f^k \cup \boldsymbol{t}_g^k|}, \tag{2}$$

where $\phi$ stands for the Jaccard similarity measure between the two subsets (Sunilkumar and Shaji, 2019). To make the similarity metric more robust, we take multiple percentage values of $k$ and plot a similarity curve of $\phi$ against varying $k$. This **Area Under the Similarity Curve (AUSC)** captures the model behavior under multiple hypotheses. AUSC is a more robust measurement of the model's indirect bias. The similarity curve also allows us to choose an optimum value of $k$ to select the most important tokens in model explanations.

The AUSC functions as a quantitative metric for assessing indirect bias present within a given text data denoted as $\boldsymbol{x}$. This metric primarily targets the identification of indirect bias at the sentence level. Nevertheless, the application scope of AUSC extends beyond individual sentences, allowing for the calculation of bias across the entire dataset. This process involves taking the AUSC values from each sentence and then calculating their average, which gives an overall measure of indirect bias in $f$ w.r.t. the entire dataset.

## 6 Indirect Bias Mitigation (IBM)

In this section, we propose a novel Indirect Bias Mitigation (IBM) algorithm to guarantee fairness in model explanations. The goal of our mitigator is to minimize the influence of protected attribute $s$ for a given model $f : \boldsymbol{x} \rightarrow y$ that is trained on labeled text data $(\boldsymbol{x}, y)$. The underlying hypothesis posits that during the training phase, $f$ picks up signals from the context tokens $\boldsymbol{t}^s$ associated with the protected attributes $s$, consequently leading to biased predictions $\hat{y}$. To mitigate such indirect bias in model explanations, we design a similarity-based regularization term $R$ to constrain the model to only rely on the key prediction centric tokens $\boldsymbol{t}^y$ but not the sensitive context tokens $\boldsymbol{t}^s$.

To obtain $R$, first, we need a pre-trained helper model $g : \boldsymbol{x} \rightarrow s$ (same as the one from IBD). During the training of our $f$ model, we take the attention matrix $A_f$ from model $f$ and the attention matrix $A_g$ from $g$ model corresponding to the same samples to calculate the cosine similarity between these two matrices using Equation 3.

$$R = (cos(A_f, A_g))^2 \,. \tag{3}$$

A greater term $R$ indicates the model $f$ relies on the sensitive context tokens $\boldsymbol{t}^s$ similarly to $g$. The preference for cosine similarity over Jaccard similarity is attributed to its differentiable nature, which is conducive to gradient-based optimization.

To achieve no indirect bias in model explanation, the model $f$ is trained with the total loss function $\mathcal{L}$ in Equation 4, where we add the similarity regularization term $R$ to the cross-entropy $CE(f(\boldsymbol{x}), y)$.

$$\mathcal{L} = CE(f(\boldsymbol{x}), y) + \lambda R, \tag{4}$$

where $\lambda$ is a hyper-parameter that controls the trade-off for fair explanations.

Our similarity regularization only aims to remove indirect bias in model explanations. It cannot guarantee the prediction outcome fairness mentioned in Section 3.1, because the layers after self-attention in the transformer-based models may still exploit the bias in the training data. In practice, it is better to complement direct bias mitigation for traditional outcome fairness with indirect bias mitigation in model explanation. In our evaluation, we show that our indirect bias mitigation is compatible with pre-process mitigation of resample (Kamiran and Calders, 2011) to and the most popular in-process mitigation for prediction outcome fairness - adversarial debiasing (AD) (Zhang et al., 2018), thus simultaneously achieving both demographic parity (or equal opportunity) in predictions and no indirect bias in model explanations.

## 7 Experiment

In this section, we evaluate our proposed Indirect Bias Discovery (**IBD**) and Indirect Bias Mitigation (**IBM**) algorithms on sentiment analysis, toxicity detection, and hate speech detection datasets. Through case studies, we also demonstrate the significance of indirect bias in model explanations and the advantage of mitigating indirect bias.

### 7.1 Datasets

The **Jigsaw Unintended Bias in Toxicity Dataset** (cjadams et al., 2019) is an archive of approximately 2 million public comments, was released at the end of 2017 following the shutdown of the Civil Comments platform. It was labeled for both the toxicity of the comments and the presence of several protected attributes. A targeted subset of this dataset, labeled specifically for toxicity towards male and female identities, comprised 21,000 records. Within this subset, 13,000 records were associated with male identities and 8,000 with male identities. The comments were classified based on toxicity levels, with 10,490 identified as toxic and 10,510 as non-toxic. The dataset has a risk difference of ∼20%, where the ratio of toxic comments towards females is higher.

The **Amazon Books Review Dataset**[1], contains feedback from 3 million users on 212,404 unique books. Using a gender inferencing model, a subset of 16,927 users (9,105 male users and 7,822 female users) was identified with high confidence based on common male and female names. This results in a subset of 33,600 reviews (16,965 positive reviews and 16,635 negative reviews), where those rated with 4 or 5 stars were classified as positive and 1-star reviews as negative. The dataset has a risk difference of ∼20%, where female users make more positive reviews. The protected attribute in this dataset is the review author's (inferred) gender. Most reviews do not include a gender self-identification token in them.

The **Measuring Hate Speech Corpus** (Sachdeva et al., 2022) comprises 50,070 social media comments, annotated by 11,143 Amazon Mechanical Turk contributors to assess hate speech through the lens of annotator perspectives, utilizing faceted Rasch measurement theory (RMT). A specific subset of this dataset, containing 27,818 comments aimed at detecting hate speech, includes 11,418 comments identified as hate speech and 16,400 as non-hate, with a focus on racial commentary—7,353 on targeting the white race and 20,460 on the black race. This subset exhibits a ∼20% higher True Positive Rate (TPR) gap for detecting hate speech against the black race.

All the datasets are split into 82% training, 8% validation, and 10% testing.

### 7.2 Metrics

We use **Accuracy** to evaluate the classification utility performance.

For prediction outcome fairness, we use Risk Difference **(RD)** to evaluate the demographic parity in model predictions for the Jigsaw dataset and the Amazon review dataset, where $RD = P(\hat{y} = 1|s = u) - P(\hat{y} = 1|s = v)$, and for the hate speech dataset, we use True Positive Rate **(TPR)** gap to evaluate equality of opportunity in model's prediction, where $TPR_{gap} = P(\hat{y} = 1|s = u, y = 1) - P(\hat{y} = 1|s = v, y = 1)$. A low RD and TPR Gap indicate fairness in terms of demographic parity and equality of opportunity respectively in the model predictions.

We use aggregated attention for model explanations and evaluate the indirect bias in model explanations using our proposed metric - Area Under Similarity Curve **(AUSC)**, which is based on the Jaccard similarity defined in Section 5. A higher value of AUSC indicates high indirect bias in the model's local explanations, where the model over-exploits sensitive context tokens in its decision-making process. In addition, we further examine the model explanations with the similarity curve (defined in Section 5). A curve below the diagonal line indicates no bias in model explanations.

To evaluate the faithfulness of our aggregator we use **comprehensiveness** $= m(x) - m(x/r)$ and **sufficiency** $= m(x) - m(r)$ (DeYoung et al., 2020) as shown in Algorithm 1, where $m(x)$ is the original prediction on $x$ of a model for a class and

[1] Amazon Books Reviews Dataset

$r$ is the rationale based on the aggregation. In Appendix A.1 we show a detailed process of extracting the most faithful combination of aggregated attention for both model $f$ and $g$. In our experiment, the combination of $(sum, sum, sum)$ for $(head\_op, layer\_op, token\_op)$ yields the best faithfulness scores.

### 7.3 Models

There is no previous work on indirect bias mitigation on model explanations. We compare our indirect bias mitigation method with some mitigation methods that focus on achieving demographic parity and equality of opportunity in predictions.

The **Vanilla Model** is a transformer-based model, we use is DistilBert (Sanh et al., 2019) and Bert (Devlin et al., 2018) with no fairness mechanism built in.

**Resampling** (Kamiran and Calders, 2011) is preprocessing mitigation, which resamples the biased dataset to get an unbiased dataset with a close to 0 risk difference. The sampled unbiased dataset is then used for the model training.

**Dropout** (Webster et al., 2021) serves as a technique to disrupt the model's ability to directly correlate protected attributes with its predictions. By intentionally increasing the Dropout rate during training, the method aims to prevent the model from overly relying on these protected attributes, a practice that can lead to overfitting. We set the Dropout rate high as 30%.

**Adversarial Debiasing (AD)** (Zhang et al., 2018) is an in-processing mitigation, which uses adversarial learning to remove the correlation between the predicted outcome and the protected attribute, i.e., achieving demographic parity (or equality of opportunity if conditioned on $y = 1$).

**Controlling Bias Exposure (CBE)** (He et al., 2022) is another in-processing mitigation technique. This method leverages an auxiliary model designed to predict a protected attribute. It utilizes the negative log-likelihood derived from this prediction as a debiasing mechanism defined as energy-based constraint. This constraint effectively regulates the significance of biased tokens, thereby controlling their influence on the model's output. We compare with it for indirect bias mitigation.

For both AD and CBE, we evaluate whether mitigation for demographic parity (or equality opportunity) and bias exposure can also lead to fairness in model explanations.

Our proposed method is to add similarity regularization for indirect bias mitigation, **IBM** on top of models that can achieve prediction outcome fairness. The helper model $g$ is trained on the same training data.

### 7.4 Performance Comparison

Due to limited space, our main results show models with the DistilBert base. The result of Bert-based models is shown in Appendix A.3. The models are evaluated on the Jigsaw and Amazon review datasets for gender bias with a high risk difference in Vanilla, and on the hate speech dataset for racial bias with a high TPR Gap in Vanilla.

#### 7.4.1 Prediction Outcome Fairness

**Demographic Parity.** In Table 1, for indirect gender bias datasets, as expected, the Vanilla model, Resampling, and Dropout cannot achieve a low risk difference in the prediction on testing data. AD, AD+CBE, and AD+IBM (Ours) achieve low risk differences through adversarial learning.

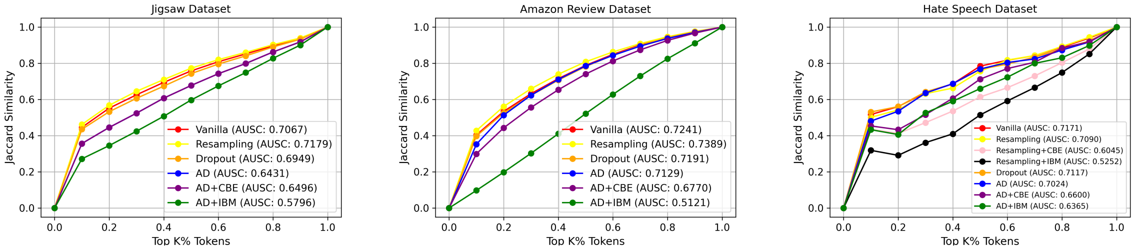
**Equality of Opportunity.** In Table 2, for the indirect racial bias dataset, Dropout achieves a slightly lower TPR Gap than Vanilla. Resampling can achieve a very low TPR Gap in the prediction. CBE and IBM can achieve low TPR Gap when paired with either Resampling or AD.

#### 7.4.2 Indirect Bias Discovery and Mitigation

**Area Under Similarity Curve (AUSC).** In both Table 1 and 2 the results for AUSC demonstrate the effectiveness of our Indirect Bias Discovery (IBD) algorithm in measuring indirect bias in model explanations across three datasets. The Vanilla model, along with Resampling, Dropout, and AD, show high AUSC scores, indicating their explanations contain indirect bias regarding the protected attribute. There is a slight correlation between prediction outcome fairness and AUSC for these models with unconstrained model attention. The only exception is that Resampling has low TPR Gap on the Hate Speech dataset but still has a high AUSC score. This is because the decision-making process is still biased for each individual record. In Table 1 for gender bias in both datasets we first see some hints of low AUSC in AD+CBE as CBE aims to control the exposure of sensitive attribute-related information, validating our AUSC metric's utility. Yet, AD+CBE cannot fully eliminate indirect bias. Our Indirect Bias Mitigation (IBM) algorithm, through similarity regularization, ensures learning from dif-

| Model | Jigsaw Dataset | | | Amazon Review Dataset | | |
|---|---|---|---|---|---|---|
| | Accuracy | RD | AUSC | Accuracy | RD | AUSC |
| Vanilla | 0.8471 | 0.2042 | 0.7067 | 0.9208 | 0.1868 | 0.7241 |
| Resampling | 0.8266 | 0.1452 | 0.7179 | 0.9169 | 0.1813 | 0.7389 |
| Dropout | 0.8433 | 0.2084 | 0.6949 | 0.9179 | 0.1869 | 0.7191 |
| AD | 0.7933 | 0.0698 | 0.6431 | 0.7477 | 0.0796 | 0.7129 |
| AD + CBE | 0.8009 | 0.0588 | 0.6496 | 0.7193 | 0.0794 | 0.6770 |
| **AD + IBM** | 0.8004 | 0.0552 | 0.5796 | 0.7254 | 0.0810 | 0.5121 |

Table 1: Model Performance on Jigsaw and Amazon Review Datasets



(a) Jigsaw Dataset

(b) Amazon Review Dataset

(c) Hate Speech Dataset

Figure 3: Similarity Curve Comparison

| Model | Accuracy | TPR Gap | AUSC |
|---|---|---|---|
| Vanilla | 0.9448 | 0.1811 | 0.7171 |
| Resampling | 0.9400 | 0.0163 | 0.7090 |
| Dropout | 0.9388 | 0.1061 | 0.7117 |
| AD | 0.9160 | 0.0357 | 0.7024 |
| Resampling + CBE | 0.9248 | 0.0531 | 0.6045 |
| AD + CBE | 0.8136 | 0.0495 | 0.6600 |
| **Resampling + IBM** | 0.9164 | 0.0381 | 0.5252 |
| **AD + IBM** | 0.8749 | 0.0502 | 0.6365 |

Table 2: Model Performance on Hate Speech Dataset

ferent patterns than those from the gender inference (helper) models. Our model explanation shows low AUSC - 0.5796 for the Jigsaw dataset and 0.5121 AUSC for the Amazon review dataset, indicating low indirect bias, i.e., the model only focuses on ground-truth-centric tokens. In Table 2 for racial bias in hate speech dataset AD+CBE, AD+IBM can achieve low AUSC of 0.6600 and 0.6365 respectively. Since Resampling has a low TPR Gap, we add CBE or IBM to mitigate indirect bias as well, which results in low AUSC - Resampling+CBE has 0.6045 AUSC and Resampling+IBM has 0.5252 AUSC. In both Resampling and AD, IBM beats CBE in terms of AUSC.

**Similarity Curve.** We can further compare the model explanation using the similarity curve. Figure 3 shows the similarity curve for each model on the three datasets, respectively. For every dataset, the Vanilla Model curve (red), the Resampling curve (yellow), the Dropout model curve (orange), and the AD curve (blue) are close to each other. The AD+CBE curve (purple) is slightly under the others. However, all five of them have a clear arch, which indicates high similarity and high indirect bias. The Resampling+CBE curve (pink), Resampling+IBM curve (black), and AD+IBM curve (green) are close to the diagonal line, which meets the goal of no indirect bias in model explanations.

### 7.4.3 Trade-off Analysis

**Trade-off Comparison.** We know there is a utility trade-off for prediction outcome fairness in machine learning (Liu and Vicente, 2022). For all three datasets, the accuracy difference between the Vanilla-biased model and AD unbiased one indicates the trade-off to achieve prediction outcome fairness (demographic parity or equal opportunity). The trade-off is 0.05, 0.17, and 0.03 for the Jigsaw, Amazon Review, and hate speech datasets, respectively. The Amazon review dataset incurs the largest accuracy drop due to the absence of the sensitive token in most texts. It is more challenging when the sensitive context is subtle. This confirms our motivation to mitigate NLP bias beyond direct bias. On the hate speech dataset, Resampling achieves equal opportunity with a utility trade-off of 0.0048. Both CBE and IBM further mitigate indirect bias in model explanations, which incurs an additional utility trade-off for fair explanations on top of AD (or Resampling). On comparison of this additional utility trade-off, AD, AD+CBE, and AD+IBM are similar in the prediction outcome fairness metrics. On the Jigsaw dataset, the additional

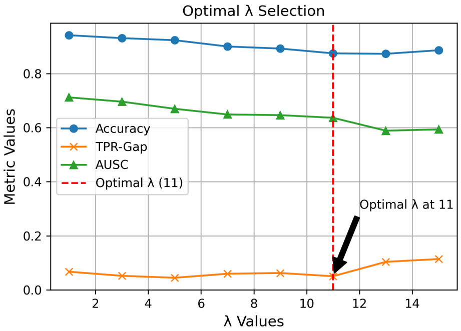


Figure 4: Sensitivity analysis of $\lambda$ on the hate speech dataset (trade-off between utility and fair explanations)

trade-off for fair explanation is almost nothing, but the indirect gender bias by AUSC is the lowest for AD+IBM. On the Amazon Review dataset, CBE and IBM have additional trade-offs of 0.0284 and 0.0223, respectively, and AUSC for AD+IBM is significantly lower. On the hate speech dataset, Resampling+CBD, AD+CBE, Resampling+IBM, and AD+IBM have additional trade-offs of 0.0152, 0.1024, 0.0236 and 0.0411, respectively. CBE regulates the model on exposure to biased tokens. It can mitigate indirect bias and reduce AUSC. Our proposed AD+IBM can achieve very low AUSC with a relatively small additional utility trade-off, i.e. AD+IBM is more effective and efficient at mitigating indirect bias in model explanations.

**Sensitivity Analysis.** In IBM, the hyperparameter $\lambda$ in Equation 4 controls the additional utility trade-off for fair explanations. Figure 4 presents a sensitivity analysis for the hyperparameter $\lambda$ for AD+IBM on the hate speech dataset for example. It illustrates that as the value of $\lambda$ escalates, there is a discernible decline in AUSC (green), at the cost of reduced accuracy (blue) and keeping TPR Gap (orange) around 0.05. The balance between AUSC, TPR Gap, and accuracy is optimized at $\lambda = 11$, which is selected as the preferable trade-off.

### 7.5 Case Analysis

To further showcase the significance of indirect bias and the advantage in its mitigation, we also conduct case analysis to directly compare different model explanations on individual examples. Figure 5 shows the explanations provided by different models of an example from the Jigsaw dataset. Due to limited space, more model explanations on other datasets are in the Appendix A.4.

Figure 5 is a toxic comment towards males from the Jigsaw dataset. All models except for AD and AD+CBE correctly predicted the toxicity. The explanations from Vanilla, Resampling, and Dropout are "men", "dominance", "priesthood", "jealous", and "fertility". They heavily overlap with the helper model's explanation for gender prediction ("men", "preisthood", and "female"). The explanation from our AD+IBM model relies on "dominance", "jealous", and "fertility", which is a gender-neutral toxicity logic. AD and AD+CBE try to put less attention on "men" and "female", but the model failed to find the toxicity logic and made the wrong prediction. We can also discover the indirect bias from these individual explanations through AUSC. Vanilla, Resampling, Dropout, AD, and AD+CBE have AUSC 0.6464, 0.6297, 0.6097, 0.5644, and 0.5226, respectively. Our AD+IBM only has 0.5030, which has the lowest indirect bias.

Our other case studies in the Appendix also shows that the other models have more similarities with the helper model while IBM focuses more on the sentiment-related content. Our AUSC score for the individual record is consistently low.

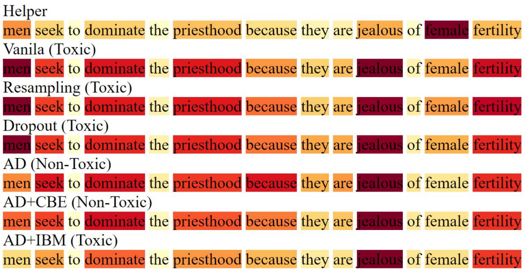


Figure 5: All model explanations on an example case from the Jigsaw dataset

## 8 Conclusion

In this work, we study indirect bias in NLP models, a phenomenon less explored but as significant as direct bias. Our contributions include defining direct versus indirect bias, introducing a new framework for quantitatively evaluating indirect bias in transformer models using their in-built self-attention matrix, and proposing a mitigation algorithm to ensure fairness in transformer models by leveraging attention explanations. Our evaluation shows the significance and challenging nature of indirect bias in model explanations, and the effectiveness of our proposed discovery and mitigation algorithms. These efforts represent a critical step towards achieving fairness and equity in NLP applications, addressing current research gaps, and guiding future ethical AI development.

## 9 Limitations

There is no publicly available dataset designed to study indirect bias. For the experiment evaluation, it is challenging to identify the ground truth-sensitive context. The current evaluation of the data we have is not enough to showcase the full spectrum of indirect bias. Our methodology heavily relies on a helper model to infer sensitive attributes. The quality of the helper model hinders the performance of our bias discovery and mitigation algorithm. The need for a helper model also slows down the runtime efficiency. In future work, we will develop a method only utilizing the target model's explanations.

## 10 Ethical Considerations

This study aims to improve NLP technology to achieve equity for all under-served communities. We want to broaden the scope of NLP fairness. Developing fair and explainable NLP models can free technology from inheriting historical bias in real-world data. Due to the limited options on datasets, we conducted the experiment with a simplified binary setting. The proposed technology is designed to comply with non-binary identities and multi-ethnicity. We hope this project raises awareness of the influence of unintentional bias from NLP models. It is a community effort to develop and advocate open-source, transparent, fair, accountable, and explainable NLP models.

## Acknowledgements

This work was supported in part by the UNC Charlotte startup fund.

## References

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022. Entropy-based attention regularization frees unintended bias mitigation from lists. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1105–1119. Association for Computational Linguistics.

Rajas Bansal. 2022. A survey on bias and fairness in natural language processing. *CoRR*, abs/2204.09591.

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias.

Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. 2022. Re-contextualizing fairness in nlp: The case of india.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4349–4357.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard S. Zemel. 2019. Understanding the origins of bias in word embeddings. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 803–811. PMLR.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791.

cjadams, Daniel Borkan, inversion, Jeffrey Sorensen, Lucas Dixon, Lucy Vasserman, and nithum. 2019. Jigsaw unintended bias in toxicity classification.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Trans. Assoc. Comput. Linguistics*, 9:1249–1267.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc Viet Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2978–2988. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Mark Diaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–14, New York, NY, USA. Association for Computing Machinery.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-dimensional gender bias classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 314–331. Association for Computational Linguistics.

Brian d'Alessandro, Cathy O'Neil, and Tom LaGatta. 2017. Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big Data*, 5(2):120–134.

Zexue He, Yu Wang, Julian McAuley, and Bodhisattwa Prasad Majumder. 2022. Controlling bias exposure for fair interpretable predictions. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5491–5501. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Faisal Kamiran and Toon Calders. 2011. Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.*, 33(1):1–33.

Svetlana Kiritchenko and Saif M. Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, pages 43–53. Association for Computational Linguistics.

Gaël Letarte, Frédérik Paradis, Philippe Giguère, and François Laviolette. 2018. Importance of self-attention for sentiment analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 267–275, Brussels, Belgium. Association for Computational Linguistics.

Suyun Liu and Luis Nunes Vicente. 2022. Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach.

Adian Liusie, Vatsal Raina, Vyas Raina, and Mark J. F. Gales. 2022. Analyzing biases to spurious correlations in text classification tasks. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2022 - Volume 2: Short Papers, Online only, November 20-23, 2022*, pages 78–84. Association for Computational Linguistics.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

Ninareh Mehrabi, Umang Gupta, Fred Morstatter, Greg Ver Steeg, and Aram Galstyan. 2022. Attributing fair decisions with attention interventions. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*. Association for Computational Linguistics.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35.

Nikolaos Mylonas, Ioannis Mollas, and Grigorios Tsoumakas. 2022. An attention matrix for every decision: Faithfulness-based arbitration among multiple attention-based interpretations of transformers in text classification.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton. 2019. Learning to deceive with attention-based explanations. *arXiv preprint arXiv:1909.07913*.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Leonid Schwenke and Martin Atzmueller. 2021. Show me what you're looking for visualizing abstracted transformer attention for enhancing their local interpretability on time series data. *The International FLAIRS Conference Proceedings*, 34.

P Sunilkumar and Athira P Shaji. 2019. A survey on semantic similarity. In *2019 International Conference on Advances in Computing, Communication and Control (ICAC3)*, pages 1–8. IEEE.

Supreme Court of the United States. 1971. Griggs v. duke power co. 401 U.S. 424. March 8.

Mike Thelwall. 2018. Gender bias in sentiment analysis. *Online Inf. Rev.*, 42(1):45–57.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Pranav Narayanan Venkit and Shomir Wilson. 2021. Identification of bias against people with disabilities in sentiment analysis and toxicity detection models.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Tianlu Wang, Rohit Sridhar, Diyi Yang, and Xuezhi Wang. 2022. Identifying and mitigating spurious correlations for improving robustness in NLP models. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1719–1729. Association for Computational Linguistics.

Yaushian Wang, Hung-Yi Lee, and Yun-Nung Chen. 2019. Tree transformer: Integrating tree structures into self-attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1061–1070, Hong Kong, China. Association for Computational Linguistics.

CS Webster, S Taylor, C Thomas, and JM Weller. 2022. Social bias, discrimination and inequity in healthcare: mechanisms, implications and recommendations. *BJA Educ*, 22(4):131–137.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2021. Measuring and reducing gendered correlations in pre-trained models.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media, SocialNLP@ACL 2020, Online, July 10, 2020*, pages 7–14. Association for Computational Linguistics.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018*, pages 335–340. ACM.

# A Appendix

## A.1 Implementation Details

For the Jigsaw and hate speech datasets, we utilize batch sizes of 32 and set the maximum token length to 128. In contrast, for the Amazon Review dataset, we opt for a batch size of 40 with a maximum token length of 256.

The $f$ and $g$ models are based on the uncased base versions of BERT and DistilBERT sequence classifiers from Huggingface, featuring 12 layers. These models undergo training over 5 epochs with a learning rate of $10^{-5}$ employing the AdamW optimizer. We implement a variant of these models for the Dropout configuration, maintaining the architecture while increasing the Dropout rate to 30%. For adversarial training, the last hidden state of model $f$ is input into the adversary model, which is a straightforward feed-forward network with two hidden layers comprising 512 and 128 units, respectively, and employs the ReLU activation function. The adversary model has a learning rate of $10^{-4}$ and utilizes cross-entropy loss for its output. In the case of the CBE model, we derive attention from both $f$ and $g$ models to compute the energy using negative log-likelihood. Lastly, our approach excludes attention from alphanumeric, punctuation, and stop-word tokens in both $f$ and $g$ models, and calculates the cosine similarity between the remaining tokens' attention.

## A.2 Faithfulness Evaluation

Here we provide the process of extracting the most faithful combination of aggregated attention as shown in Algorithm 1 for both model $f$ and $g$ using the Jigsaw dataset for example. The process begins by evaluating the $g$ model's comprehensiveness and sufficiency across various aggregation strategies—namely summation, multiplication, averaging, and maximization—applied at different structural levels: head, layer, and matrix. This evaluation involves examining the top 20%, 30%, and 40% of tokens, as outlined in Table 3a. Subsequently, we select the aggregation combination that yields the highest scores in comprehensiveness and lowest in sufficiency for the top $k\%$ of tokens - for the Jigsaw dataset, we take the $sum, sum, sum$ combination. The chosen combination and token percentage are then applied to model $f$. The analysis of the $f$ model includes the performance of the Vanilla model and our model (AD+IBM), with findings presented in Table 3b. This approach allows us to systematically determine the aggregation technique that most effectively maintains the faithfulness of the attention mechanism in mitigating bias.

## A.3 Model Performance with Bert-Based Models

Table 4 shows the result of our evaluation of Bert-based models.

**Demographic Parity.** For both datasets, as expected, neither Resampling nor Dropout can achieve low risk difference in the prediction on testing data. AD, AD+CBE and AD+IBM can achieve low risk differences through adversarial learning.

**Indirect Bias Discovery and Mitigation.** For both datasets, the other models all have high AUSC scores (above 0.7), which means their explanations have indirect bias w.r.t. the protected attribute. For our Indirect Bias Mitigation (IBM) algorithm, the similarity regularization makes sure the model learns different patterns from the gender inference (helper) model. Our model explanation has a close to 0.5 AUSC, indicating low indirect bias, i.e., the model only focuses on the ground-truth-centric tokens.

## A.4 Additional Case Analysis

Figure 6 is a negative review by a female author from the Amazon Review dataset. All models correctly predicted the negative sentiment. The explanations from Vanilla and other baselines have more similarities with the helper model to detect female gender. For our AD+IBM model, the explanation focuses more on the sentiment-related content and the attention is spread out evenly. The indirect bias discovered in the AUSC score for ours is only 0.5594 compared to other models having around 0.75.

Figure 7 is case study from the hate speech dataset. All models correctly predicted hate speech. The explanations from Vanilla, Resampling, Dropout, AD, and AD+CBE put more emphasis on the word "ni**as", which is a keyword for the helper model. For our AD+IBM model, the explanation focuses more on the hate related word (e.g., "bitch", "hate", "fuck", etc.). This means our mitigator avoids potentially sensitive context and focuses only on ground-truth-centric tokens. The indirect bias discovered in the AUSC score for Vanilla, Resampling, Dropout, AD, and AD+CBE is 0.7495, 0.7346, 0.7112, 0.7200, and 0.7094, respectively. Resampling+CBE, Resampling+IBM,

| $g$ model | sum,sum,sum | | mean,max,mul | | sum,mean,mul | |
|---|---|---|---|---|---|---|
| Accuracy = 0.9714 | Com | Suff | Com | Suff | Com | Suff |
| 20% | 0.3433 | 0.0100 | 0.2400 | 0.0885 | 0.3371 | 0.0133 |
| 30% | 0.3504 | 0.0090 | 0.2895 | 0.0419 | 0.3461 | 0.0104 |
| 40% | 0.3538 | 0.0085 | 0.3180 | 0.0200 | 0.3504 | 0.0095 |

(a) Faithfulness of the $g$ model

| Model | Accuracy | Com | Suff |
|---|---|---|---|
| Vanilla | 0.8433 | 0.2295 | 0.0490 |
| AD+IBM | 0.7614 | 0.1257 | 0.0285 |

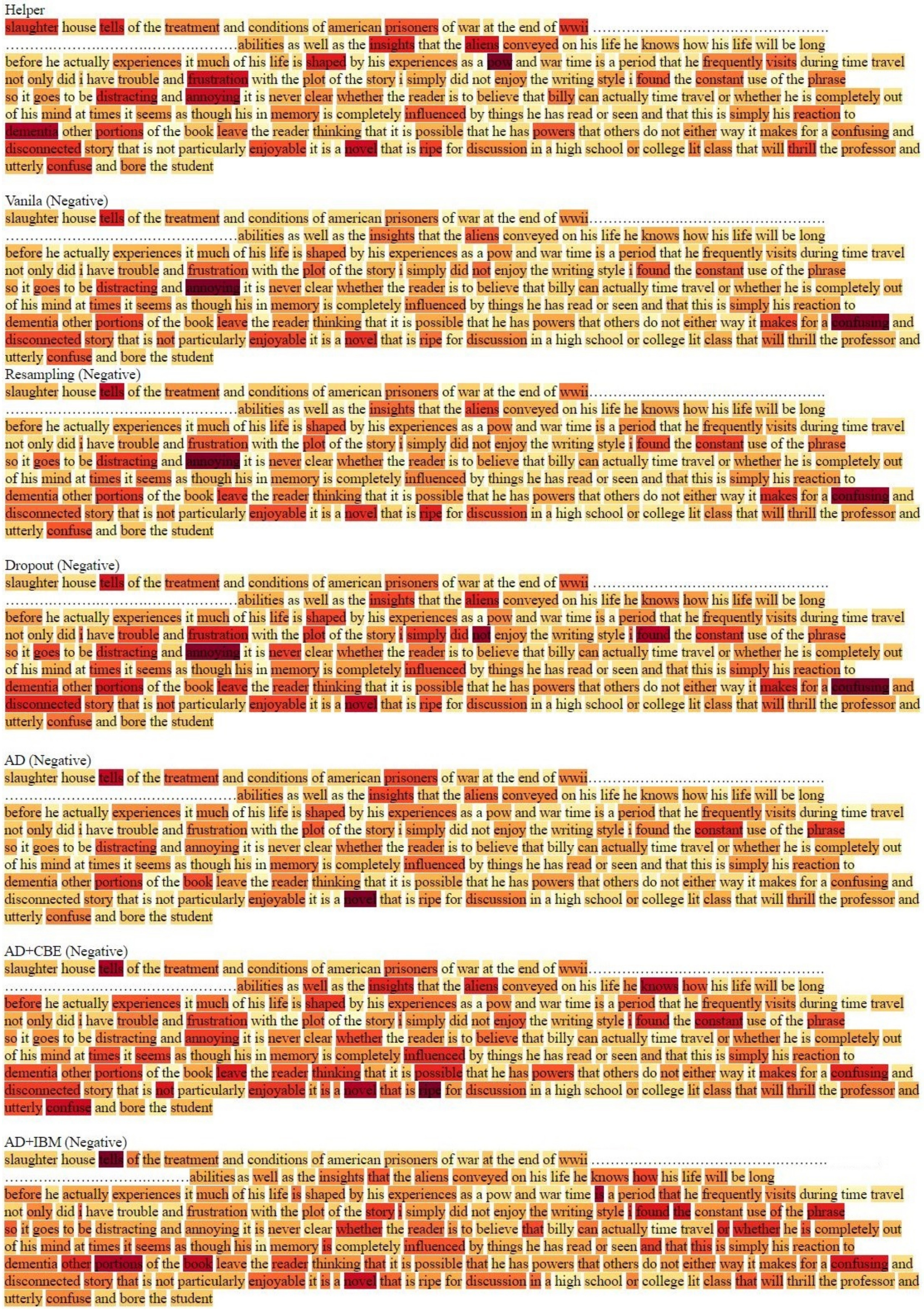
(b) Faithfulness of the $f$ models at 30% on sum, sum, sum combination

Table 3: Faithfulness evaluation $f$ and $g$ model on Jigsaw dataset

| Model (Bert) | Jigsaw Dataset (Gender) | | | Amazon Review Dataset (Gender) | | |
|---|---|---|---|---|---|---|
| | Accuracy | RD | AUSC | Accuracy | RD | AUSC |
| Vanilla | 0.8433 | 0.1928 | 0.7406 | 0.9362 | 0.1941 | 0.7752 |
| Resampling | 0.8487 | 0.1635 | 0.7478 | 0.9297 | 0.1849 | 0.7685 |
| Dropout | 0.8357 | 0.2250 | 0.7265 | 0.9032 | 0.1776 | 0.7660 |
| AD | 0.7928 | 0.0694 | 0.7275 | 0.7627 | 0.0741 | 0.7274 |
| AD + CBE | 0.8004 | 0.0533 | 0.7119 | 0.7092 | 0.0673 | 0.7634 |
| **AD + IBM** | 0.7615 | 0.0332 | 0.5906 | 0.7242 | 0.0823 | 0.5540 |

Table 4: Model Performance on Jigsaw and Amazon Review Datasets with Bert-based models

and AD+IBM have AUSC of 0.6157, 0.4992, and 0.6107, respectively. They are better at mitigating indirect bias.



Helper
slaughter house tells of the treatment and conditions of american prisoners of war at the end of wwii ………………………………………………
………………………………………………abilities as well as the insights that the aliens conveyed on his life he knows how his life will be long
before he actually experiences it much of his life is shaped by his experiences as a pow and war time is a period that he frequently visits during time travel
not only did i have trouble and frustration with the plot of the story i simply did not enjoy the writing style i found the constant use of the phrase
so it goes to be distracting and annoying it is never clear whether the reader is to believe that billy can actually time travel or whether he is completely out
of his mind at times it seems as though his in memory is completely influenced by things he has read or seen and that this is simply his reaction to
dementia other portions of the book leave the reader thinking that it is possible that he has powers that others do not either way it makes for a confusing and
disconnected story that is not particularly enjoyable it is a novel that is ripe for discussion in a high school or college lit class that will thrill the professor and
utterly confuse and bore the student

Vanila (Negative)
slaughter house tells of the treatment and conditions of american prisoners of war at the end of wwii………………………………………………
………………………………………………abilities as well as the insights that the aliens conveyed on his life he knows how his life will be long
before he actually experiences it much of his life is shaped by his experiences as a pow and war time is a period that he frequently visits during time travel
not only did i have trouble and frustration with the plot of the story i simply did not enjoy the writing style i found the constant use of the phrase
so it goes to be distracting and annoying it is never clear whether the reader is to believe that billy can actually time travel or whether he is completely out
of his mind at times it seems as though his in memory is completely influenced by things he has read or seen and that this is simply his reaction to
dementia other portions of the book leave the reader thinking that it is possible that he has powers that others do not either way it makes for a confusing and
disconnected story that is not particularly enjoyable it is a novel that is ripe for discussion in a high school or college lit class that will thrill the professor and
utterly confuse and bore the student

Resampling (Negative)
slaughter house tells of the treatment and conditions of american prisoners of war at the end of wwii………………………………………………
………………………………………………abilities as well as the insights that the aliens conveyed on his life he knows how his life will be long
before he actually experiences it much of his life is shaped by his experiences as a pow and war time is a period that he frequently visits during time travel
not only did i have trouble and frustration with the plot of the story i simply did not enjoy the writing style i found the constant use of the phrase
so it goes to be distracting and annoying it is never clear whether the reader is to believe that billy can actually time travel or whether he is completely out
of his mind at times it seems as though his in memory is completely influenced by things he has read or seen and that this is simply his reaction to
dementia other portions of the book leave the reader thinking that it is possible that he has powers that others do not either way it makes for a confusing and
disconnected story that is not particularly enjoyable it is a novel that is ripe for discussion in a high school or college lit class that will thrill the professor and
utterly confuse and bore the student

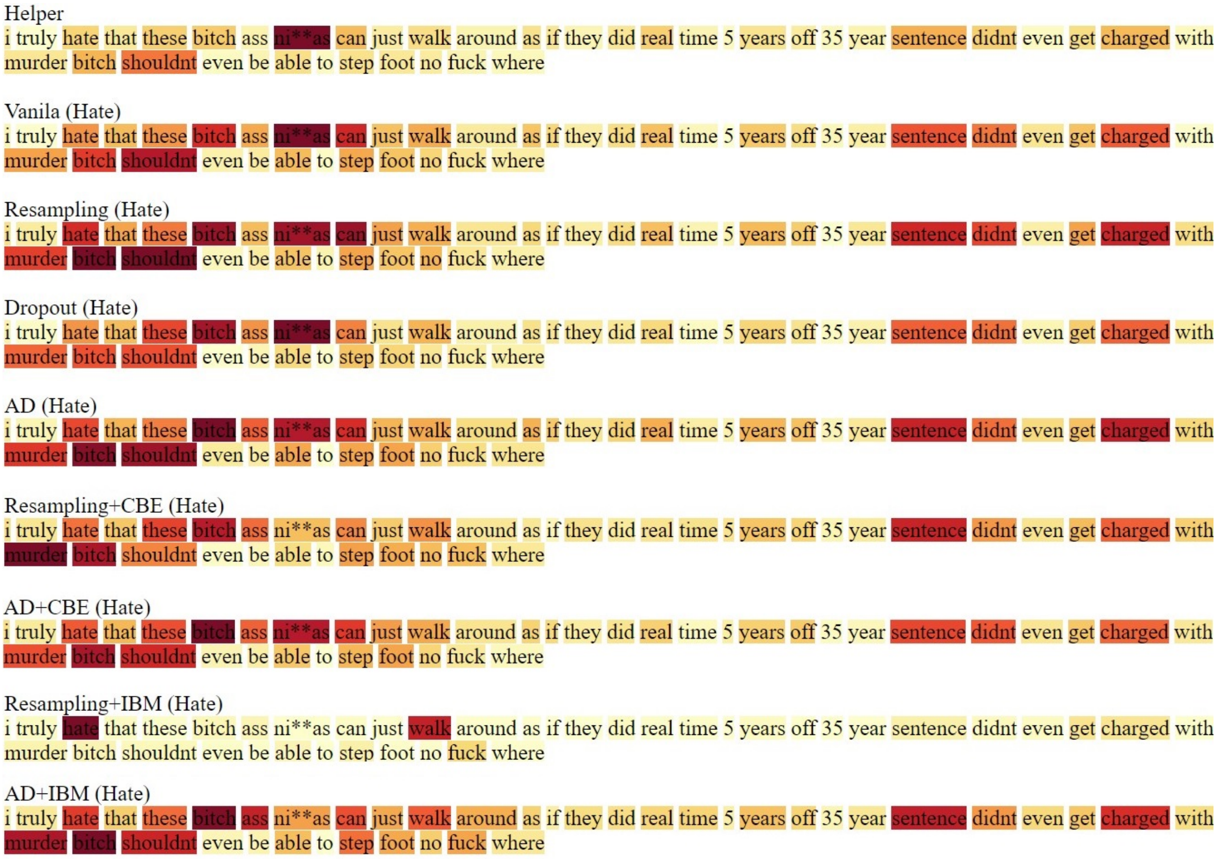
Dropout (Negative)
slaughter house tells of the treatment and conditions of american prisoners of war at the end of wwii ………………………………………………
………………………………………………abilities as well as the insights that the aliens conveyed on his life he knows how his life will be long
before he actually experiences it much of his life is shaped by his experiences as a pow and war time is a period that he frequently visits during time travel
not only did i have trouble and frustration with the plot of the story i simply did not enjoy the writing style i found the constant use of the phrase
so it goes to be distracting and annoying it is never clear whether the reader is to believe that billy can actually time travel or whether he is completely out
of his mind at times it seems as though his in memory is completely influenced by things he has read or seen and that this is simply his reaction to
dementia other portions of the book leave the reader thinking that it is possible that he has powers that others do not either way it makes for a confusing and
disconnected story that is not particularly enjoyable it is a novel that is ripe for discussion in a high school or college lit class that will thrill the professor and
utterly confuse and bore the student

AD (Negative)
slaughter house tells of the treatment and conditions of american prisoners of war at the end of wwii………………………………………………
………………………………………………abilities as well as the insights that the aliens conveyed on his life he knows how his life will be long
before he actually experiences it much of his life is shaped by his experiences as a pow and war time is a period that he frequently visits during time travel
not only did i have trouble and frustration with the plot of the story i simply did not enjoy the writing style i found the constant use of the phrase
so it goes to be distracting and annoying it is never clear whether the reader is to believe that billy can actually time travel or whether he is completely out
of his mind at times it seems as though his in memory is completely influenced by things he has read or seen and that this is simply his reaction to
dementia other portions of the book leave the reader thinking that it is possible that he has powers that others do not either way it makes for a confusing and
disconnected story that is not particularly enjoyable it is a novel that is ripe for discussion in a high school or college lit class that will thrill the professor and
utterly confuse and bore the student

AD+CBE (Negative)
slaughter house tells of the treatment and conditions of american prisoners of war at the end of wwii………………………………………………
………………………………………………abilities as well as the insights that the aliens conveyed on his life he knows how his life will be long
before he actually experiences it much of his life is shaped by his experiences as a pow and war time is a period that he frequently visits during time travel
not only did i have trouble and frustration with the plot of the story i simply did not enjoy the writing style i found the constant use of the phrase
so it goes to be distracting and annoying it is never clear whether the reader is to believe that billy can actually time travel or whether he is completely out
of his mind at times it seems as though his in memory is completely influenced by things he has read or seen and that this is simply his reaction to
dementia other portions of the book leave the reader thinking that it is possible that he has powers that others do not either way it makes for a confusing and
disconnected story that is not particularly enjoyable it is a novel that is ripe for discussion in a high school or college lit class that will thrill the professor and
utterly confuse and bore the student

AD+IBM (Negative)
slaughter house tells of the treatment and conditions of american prisoners of war at the end of wwii ………………………………………………
……………………………………abilities as well as the insights that the aliens conveyed on his life he knows how his life will be long
before he actually experiences it much of his life is shaped by his experiences as a pow and war time is a period that he frequently visits during time travel
not only did i have trouble and frustration with the plot of the story i simply did not enjoy the writing style i found the constant use of the phrase
so it goes to be distracting and annoying it is never clear whether the reader is to believe that billy can actually time travel or whether he is completely out
of his mind at times it seems as though his in memory is completely influenced by things he has read or seen and that this is simply his reaction to
dementia other portions of the book leave the reader thinking that it is possible that he has powers that others do not either way it makes for a confusing and
disconnected story that is not particularly enjoyable it is a novel that is ripe for discussion in a high school or college lit class that will thrill the professor and
utterly confuse and bore the student

Figure 6: All model explanations on an example case from the Amazon review dataset



Helper
i truly hate that these bitch ass ni**as can just walk around as if they did real time 5 years off 35 year sentence didnt even get charged with murder bitch shouldnt even be able to step foot no fuck where

Vanila (Hate)
i truly hate that these bitch ass ni**as can just walk around as if they did real time 5 years off 35 year sentence didnt even get charged with murder bitch shouldnt even be able to step foot no fuck where

Resampling (Hate)
i truly hate that these bitch ass ni**as can just walk around as if they did real time 5 years off 35 year sentence didnt even get charged with murder bitch shouldnt even be able to step foot no fuck where

Dropout (Hate)
i truly hate that these bitch ass ni**as can just walk around as if they did real time 5 years off 35 year sentence didnt even get charged with murder bitch shouldnt even be able to step foot no fuck where

AD (Hate)
i truly hate that these bitch ass ni**as can just walk around as if they did real time 5 years off 35 year sentence didnt even get charged with murder bitch shouldnt even be able to step foot no fuck where

Resampling+CBE (Hate)
i truly hate that these bitch ass ni**as can just walk around as if they did real time 5 years off 35 year sentence didnt even get charged with murder bitch shouldnt even be able to step foot no fuck where

AD+CBE (Hate)
i truly hate that these bitch ass ni**as can just walk around as if they did real time 5 years off 35 year sentence didnt even get charged with murder bitch shouldnt even be able to step foot no fuck where

Resampling+IBM (Hate)
i truly hate that these bitch ass ni**as can just walk around as if they did real time 5 years off 35 year sentence didnt even get charged with murder bitch shouldnt even be able to step foot no fuck where

AD+IBM (Hate)
i truly hate that these bitch ass ni**as can just walk around as if they did real time 5 years off 35 year sentence didnt even get charged with murder bitch shouldnt even be able to step foot no fuck where

Figure 7: All model explanations on an example case from the hate speech dataset